ANUARI DE FILOLOGIA. ESTUDIS DE LINGÜÍSTICA(Anu.Filol.Est.Lingüíst.)
3/2013, pp. 203-210, ISSN: 2014-1408

TRADUCCIÓN DE CONTENIDOS Y APLICACIONES CON UN SISTEMA AUTOMÁTICO INTEGRADO DE POSTEDICIÓN Y MEJORA DE LA PRODUCTIVIDAD

ÓSCAR SUAU VERDIGUIER
CEO AutomaticTrans
osuau@automatictrans.es

RESUMEN

AutomaticTrans (AT) es una empresa que desde 1994 se dedica a desarrollar y comercializar tecnología para facilitar los procesos multilingües de grandes organizaciones. Las empresas que utilizan la tecnología de AT se benefician de la organización de los procesos, los controles de calidad sistemáticos y la eficacia en las tareas de traducción y localización.

Los desarrollos de AT están orientados en tres áreas: lingüística, gestión e integración. Lingüística para facilitar con traducción automática y memorias de traducción las tareas de los traductores; gestión para organizar todo el ciclo de procesos desde el pedido hasta la publicación; integración para facilitar la puesta en marcha en organizaciones complejas de todos los procesos alineados con la tecnología ya existente.

PALABRAS CLAVE: tecnología; traducción automática; postedición; www; websites; aplicaciones; multilingüismo.

APPS AND CONTENT TRANSLATION USING AN INTEGRATED AND AUTOMATIC SISTEM TO POSTEDITING AND PRODUCTIVITY IMPROVEMENT

ABSTRACT

Since 1994, AutomaticTrans (AT) is a technical and commercial company focused on solutions for multilingual customers making easier this complex process for big organizations. The companies working with AT solutions improves their workflow organization, adds a systematic quality control and benefits of an efficient translation and localization task model.

AT products cover 3 main areas: linguistic, management and integration. Linguistic by providing machine translation and translation memories tools to translators and reviewers; management for the complete life cycle, end-to-end (invoicing to publishing) tasks; integration to make easy the immediate installation and deployment in complex companies by aligning the multilingual processes with the legacy technology of each company.

KEY WORDS: technology; machine translation; postediting; www; websites; applications; multilingual.

1. INTRODUCCIÓN

En este artículo se hace una descripción de las áreas de investigación y desarrollo en las tres áreas de actividad principal de AutomaticTrans (AT), a partir de la tecnología que se ha desarrollado en los años anteriores.

AT empezó por intentar resolver la parte lingüística con un sistema básico de traducción que empleaba memorias de traducción, patrones y sistemas de aproximación (*fuzzy*) para la traducción de textos. Estos sistemas empíricos se

Data de recepció: 02/IX/2013
Data d'acceptació: 05/X/2013

han desarrollado en las distintas evoluciones de la tecnología hacia un sistema que emplea tres mecanismos: las memorias de traducción con patrones (1994), el sistema de reglas (1999) y el sistema estadístico (2003). Cada técnica se aplica de forma sucesiva para intentar resolver las traducciones con el mejor método posible. Estas evoluciones se han conseguido en paralelo con las mejoras de rendimiento de la tecnología de base, ya que cada evolución ha precisado de una mayor potencia de computación y de arquitecturas de sistemas más combinables.

En la parte de gestión se ha trabajado en los aspectos administrativos (contaje de palabras, manejo de proyectos, soporte a las personas y control de la productividad) y técnicos (gestión de sistemas, disponibilidad y rendimiento). En este sentido los sistemas comerciales deben cumplir con las premisas de eficacia de un sistema crítico para empresas que necesitan la traducción en el día a día de sus operaciones, como si se tratara de cualquier sistema de producción de su red de sistemas.

En cuanto a la integración, los sistemas de AT están diseñados para la interoperabilidad en arquitecturas SOA (Services Oriented Architecture) que permitan la inmediata adaptación a los entornos ya existentes a los clientes. Primero para establecer tráficos de documentos y gestores de contenidos y desde 2007 para la interoperabilidad de aplicaciones, redes sociales y sistemas de posicionamiento.

2. APLICACIONES EN DESARROLLO

Los desarrollos de AT siguen evolucionando en las tres áreas para cumplir con el objetivo de cubrir las necesidades de extremo a extremo de todo el proceso. Por tanto el esfuerzo se está realizando en los aspectos que inciden en la productividad de los profesionales implicados.

2.1. Lingüística

La parte lingüística se desarrolla orientada a resolver las casuísticas personalizadas para cada cliente. A diferencia de la aproximación a la traducción universal que ya intentan resolver Google o Microsoft, AT siempre ha desarrollado sus sistemas para cubrir las necesidades de clientes concretos. Cada uno tiene un estilo propio en la redacción de originales, con uso de frases subordinadas en mayor o menor medida y con una calidad ortogramatical desigual. Para cada uno hay que atender a unas especificaciones concretas.

La aproximación personalizada, por otro lado, facilita el trabajo porque la cantidad de unidades a tratar de esta forma se limita a un volumen léxico reducido, que una vez categorizados concretan más su aplicabilidad y desambiguación. Este modelo ha sido el único que sabemos aplicar que se ha demostrado práctico para resolver los problemas de la traducción con calidad.

Como ya se ha dicho facilita la resolución de ambigüedad, polisemia y tratamiento del discurso (figuras retóricas, alegóricas y metafóricas).

La consideración de nuevas técnicas siempre está en la cabeza de los que dirigen la estrategia lingüística de AT que siguen los estudios que realizan distintos centros de investigación. Esperamos encontrar un cambio revolucionario en la forma de plantear el problema ya que las técnicas actuales no nos hacen vislumbrar una solución efectiva (productiva) al reto de la traducción automática de calidad.

AT aplica la estadística desde el año 2003 en varias etapas del trabajo de los lingüistas. Para AT esta técnica es útil en las fases de análisis de corpus y personalización de los recursos, alineación, detección de incoherencias y generación de patrones y memorias de traducción. Sin embargo, aplicada como único modo a la traducción resulta insuficiente y mantiene un techo demasiado bajo para los procesos de calidad que se necesitan en la industria lingüística.

Es cierto que la traducción estadística (SMT, Koehn *et al.* 2007) está sirviendo para que una buena cantidad de empresas nazcan dentro del sector, gracias a la facilidad de la puesta en marcha de estos sistemas utilizando aplicaciones de código abierto, la más extendida es MOSES. Pero su nacimiento no ha supuesto una evolución práctica (sí divulgativa) de la traducción automática. La pretraducción que se consigue en muchos casos con la SMT es improductiva, cuando no contraproducente ya que puede traducir con distinto sentido al que tiene el original y crear confusión. Hay que considerar aparte a clientes con problemas de traducción muy básicos (manuales simples, instrucciones de uso o websites con expresiones muy limitadas), donde sí está siendo de utilidad, pero de la misma manera que lo sería la memoria de traducción.

Sin embargo, la calidad necesaria para traducir documentos más complejos como manuales detallados, contratos (llenos de frases subordinadas), ofertas para licitaciones, memorias corporativas, informes, documentos legales, información divulgativa en general y documentos descriptivos se sigue precisando de una atenta labor de los equipos humanos de traducción.

Por estas causas, la evolución en AT se sigue centrando en el desarrollo de sistemas de reglas (Aitchison 1999) cada vez más sofisticados que permitan, sin penalizar el rendimiento (un aspecto fundamental en los desarrollos de AT) solucionar casos concretos que aparecen en los corpus específicos de los clientes. Lógicamente esto reduce la capacidad de AT para desarrollar nuevos pares de idiomas, que si se dispone del corpus paralelo necesario se generan con rapidez en sistemas SMT. Esta línea de desarrollo también encarece el proceso, pero por otro lado permite ofrecer un nivel de calidad más productivo, pensando en la orientación a la postedición en proyectos concretos y a largo plazo. La métrica de productividad que se emplea en el sector se mide en palabras por jornada: un traductor que emplea herramientas CAT (Computer-Assisted Translation, Moré *et al.* 2004) convencionales (TRADOS, DEJAVU,

OMEGAT) tiene una producción media (o se compromete a traducir) de 2.500 a 3.500 palabras por jornada de 8 horas. Los traductores que trabajan con los sistemas de AT doblan esta productividad en pares de lenguas distantes (español a inglés, alemán, ruso, chino o viceversa) y la triplican en lenguas romances (español, portugués, francés, italiano, catalán o gallego).

Bajo el punto de vista de AT, salvo un cambio de paradigma respecto a los métodos actuales de traducción automática, con los planteamientos actuales, no sabemos desarrollar una lógica precisa que ofrezca el nivel de calidad imprescindible para una producción de las traducciones por encima de los valores antes mencionados. Es por esta razón que pese a las corrientes actuales persistimos en nuestro planteamiento de desarrollar modelos basados en sistemas híbridos que se beneficien de lo mejor de cada técnica.

Los desarrollos de AT, además de mejorar el sistema de reglas, están también orientados al direccionamiento ilimitado de memorias de traducción, gestionando grandes volúmenes, con categorización dinámica de los segmentos y para el soporte de los sistemas de control de calidad en la postedición.

Estas son las líneas de trabajo con mayor inversión en AT y donde se dedican los esfuerzos de desarrollo.

Es importante destacar que en AT los mecanismos para la creación de reglas y patrones están al alcance de un lingüista sin conocimientos de informática, únicamente debe tener las ideas claras para formular la mejor estrategia aplicable a un problema lingüístico. En este sentido, los sistemas que dan soporte a la creación de estrategias de traducción mediante reglas tienen la ventaja de poder generar ejemplos de uso de cada modelo que se rastrean en los corpus de referencia, para que el lingüista compruebe los resultados en tiempo real de cada hipótesis. Las comprobaciones las realiza el sistema de validación de reglas usando el corpus de referencia, el propio del cliente o el adaptado para temáticas específicas donde será de aplicación según la categoría en la que se sitúa cada regla. De este modo, la creación de conjuntos de reglas es más simple que en sistemas donde la programación y el desarrollo lingüístico son altamente dependientes. Este es un modelo que se diseñó así desde el principio para poder desarrollar nuevos pares con la metodología más productiva que hemos sabido concretar para personas expertas en la definición de modelos lingüísticos, sin intervención de informáticos.

La combinación de la estadística y de la lingüística (diccionarios, patrones y reglas) ofrece unos resultados suficientemente productivos para los proyectos en los que trabaja AT. Este modelo híbrido se explota siguiendo una metodología en la que el tratamiento terminológico, de cada categoría, su etiquetación, el tratamiento de palabras adyacentes y el tratamiento de los verbos permite avanzar garantizando la calidad de los automatismos.

Los ciclos de retroalimentación automática, desde los módulos de postedición, alimentan los corpus paralelos que permiten detectar incidencias que no se previeron en el desarrollo de las reglas, aportando información y

ejemplos para la mejora de los modelos, para la adición de patrones o mejora de las reglas.

La aplicación de técnicas para el manejo de grandes volúmenes de corpus (Hilbert 2013) junto con herramientas de etiquetado derivadas del procesamiento del lenguaje (Yucong Duan 2011) permiten una mejora en los sistemas orientados a la detección y extracción de ciertas partes del contenido (entidades nombradas, verbos compuestos, polisemias, ambigüedades) que facilitan el tratamiento de los problemas más críticos de los sistemas de traducción. Los avances en esta línea están orientados a la descriptiva de los casos a solucionar. En el caso de AT no prevemos desarrollos específicos en esta línea, sino mejorar los que ya están en producción en estos momentos para los cuales es fundamental el trabajo de los lingüistas.

2.2. Gestión

Para la gestión de los procesos de traducción existen una serie de módulos en la plataforma de AT orientados a la parte administrativa que no mencionaremos. Los que son de interés para el marco de este artículo son los módulos dedicados al cuidado de un aspecto crítico para una empresa de traducción: la evaluación del trabajo de los traductores, su calidad y su productividad.

Esta evaluación aprovecha información de los sistemas de postedición que emplean herramientas lingüísticas para el control de cambios, correctores ortotipográficos y gramaticales a partir de los recursos personalizados para cada proyecto. Aquí de nuevo las técnicas de procesado del lenguaje se aplican para conseguir un sistema justo de puntuación en cada trabajo que permita disponer de un cuadro de evaluación del trabajo de las personas implicadas en la traducción. También sirven para evaluar el resultado productivo real de los automatismos.

Los resultados de cada traducción se anotan en los registros de evaluación que luego son analizados por los coordinadores de cada par de lenguas y los jefes de proyecto. De la información de los registros se desprende:

- el volumen de correcciones realizadas,
- la correspondencia con la personalización del cliente (libro de estilo),
- las incoherencias de la terminología,
- el grado de acierto del sistema de traducción.

Las herramientas de corrección emplean un conjunto de recursos muy similar al empleado en la traducción con funciones específicas de utilidad en la corrección de traducciones:

- concordancias,
- conjugaciones,
- gramática.

Lógicamente también marca los errores más evidentes, detectables por la simple aplicación del diccionario para revisar los aspectos ortotipográficos y de formato.

El corrector de AT, a diferencia de los empleados en los procesadores de texto, puede marcar también las palabras polisémicas y ambiguas para que los correctores puedan ver de manera destacada estas situaciones y si su aplicación ha sido la correcta.

2.3. Integración

Desde 2007 AT ha centrado una parte importante de su actividad en el desarrollo de módulos de traducción directa de portales de Internet o Intranet para evitar la compleja gestión de portales multilingües.

Inicialmente los portales disponían de uno o dos idiomas, ahora tenemos clientes con 9 o 12 idiomas y su gestión por mecanismos convencionales o por la integración de los gestores de contenidos no son sostenibles ni económica ni organizativamente.

La solución propuesta permite realizar las siguientes acciones:

- descarga selectiva de portales mono- o multilingües,
- alineación de los contenidos multilingües,
- descarga de documentos asociados,
- tratamiento de la codificación y separación de los textos,
- análisis lexicométrico,
- evaluador de traducciones de AT comparadas con las existentes en los sitios multilingües, para normalizar reglas,
- generador de patrones de los segmentos diferentes,
- organización categorizada para los lingüistas,
- revisión por los lingüistas y generación/corrección de patrones,
- generación de reglas personalizadas,
- validación de los resultados automáticos según modelos concretos para cada cliente y proyecto.

En estos procesos entran técnicas de LT (Linguistic Technology) con herramientas especializadas para la detección de idiomas, detección de lemas, alineación, detección de categorías, evaluación y generación de patrones. En cada caso se utilizan desarrollos especializados para cada tipo de función.

Además de la recopilación, en la parte de integración se tratan también aspectos técnicos no relacionados con las LT pero que son necesarios para su aplicación. Los puntos principales que se gestionan son los relativos a la seguridad y confidencialidad, la disponibilidad y las redes y sistemas.

El conjunto de técnicas se emplea para configurar los servicios con las peculiaridades de cada proyecto. Una vez determinado el modelo de traducción contratado en cada proyecto entran en operación los módulos de la plataforma necesarios que permitirán la traducción inicial y el mantenimiento continuado

del portal de Internet y de los documentos enviados por los distintos usuarios del cliente. De este modo se establece un punto centralizado de traducción para todas las necesidades de un cliente. Al centralizar el servicio se mantiene sistemáticamente la coherencia lingüística de todos los contenidos, cosa que sin este planteamiento era difícil, si no imposible, de gestionar debido a la atomización de los servicios de traducción.

Las funciones de integración que realizan los sistemas son:

- análisis de la web original para detección de nuevos contenidos y de cambios,
- descarga de nuevas páginas o aplicaciones,
- descarga de modificaciones en los contenidos,
- pretraducción por el sistema automático (que previamente ha sido personalizado según la metodología lingüística que se ha definido en el apartado anterior),
- flujo de postedición colaborativo donde todos los proveedores están alienados en la terminología y libro de estilo acordado con el cliente,
- flujo de mejora del sistema de traducción automática a partir de las correcciones realizadas en el flujo de postedición,
- publicación automática del portal traducido, contenidos y aplicaciones,
- registro y gestión de todas las actividades implicadas,
- informas de servicio y
- control de calidad de todo el proceso

La integración de todos estos componentes en un flujo homogéneo y adaptable a cada tipo de organización, cada una con su sistema de información específico, es uno de los factores de mayor éxito de nuestra propuesta. Debido a esta facilidad la deducción en los costes de un proyecto de traducción están por encima del 30% respecto al coste total y en algunos casos llega hasta el 80%, cuando la reprogramación de las aplicaciones en varios idiomas significa gastos inasumibles por el medio convencional de traducción.

3. APLICACIONES POSIBLES

Para AT el campo de aplicación de las tecnologías lingüísticas está centrado en la traducción: traducción automática y herramientas de soporte. Su aplicación real desde hace más de 15 años nos ha representado alcanzar un nivel de producción relativamente satisfactorio, al límite de las posibilidades que sabemos gestionar en estos momentos.

Es posible que se determine un cambio de paradigma que permita avanzar en la solución automatizada, pero bajo el punto de vista de los lingüistas de AT con la formalización de las técnicas hoy conocidas va a ser muy difícil, si no imposible, llegar a un nivel superior a un coste asumible.

BIBLIOGRAFÍA

AITCHISON, J. (1999[2] [1995]), *Linguistics: An Introduction*, London, Hodder&Stoughton.

HILBERT, M. (2013), "Big Data for Development: From Information to Knowledge Societies", SSRN Scholarly Paper No.ID 2205145, Rochester, NY, Social ScienceResearch Network.

MORÉ LÓPEZ, J. y OLIVER GONZÁLEZ, A. (2004), *Traducción asistida por ordenador: programas y recursos libres y gratuitos: material bilingüe*, Climent Roca, S. (coor.), Barcelona, Planeta UOC.

KOEHN, P., HOANG, H., BIRCH A., CALLISON-BURCH, C., FEDERICO M., BERTOLDI, N., COWAN, B., SHEN, W., MORÁN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. y HERBST, E. (2007), *Moses: Kit de herramientas de código abierto para la traducción automática estadística*, ACL 2007, Sesión de demostración, Praga, República Checa.

YUCONG, D. y CRUZ, C. (2011), "Formalizing Semantic of Natural Language through Conceptualization from Existence", *International Journal of Innovation, Management and Technology* (2011) 2 (1), 37-42.